Corpus Linguistics and Eighteenth Century Collections Online (ECCO)

Tolonen, Mikko

2021

Tolonen , M , Mäkelä , E , Ijaz , A & Lahti , L 2021 , ' Corpus Linguistics and Eighteenth Century Collections Online (ECCO) ' , Research in Corpus Linguistics , vol. 9 , no. 1 , pp. 19-34 . < https://ricl.aelinco.es/index.php/ricl/article/view/161 >

http://hdl.handle.net/10138/333313

Research in

Corpus Linguistics and *Eighteenth Century Collections Online* (ECCO)

Mikko Tolonen[a] – Eetu Mäkelä[a] – Ali Ijaz[a] – Leo Lahti[b]
University of Helsinki[a] / Finland
University of Turku[b] / Finland

Abstract – *Eighteenth Century Collections Online* (ECCO) is the most comprehensive dataset available in machine-readable form for eighteenth-century printed texts. It plays a crucial role in studies of eighteenth-century language and it has vast potential for corpus linguistics. At the same time, it is an unbalanced corpus that poses a series of different problems. The aim of this paper is to offer a general overview of ECCO for corpus linguistics by analysing, for example, its publication countries and languages. We will also analyse the role of the substantial number of reprints and new editions in the data, discuss genres and the estimates of Optical Character Recognition (OCR) quality. Our conclusion is that whereas ECCO provides a valuable source for corpus linguistics, scholars need to pay attention to historical source criticism. We have highlighted key aspects that need to be taken into consideration when considering its possible uses.

Keywords – *Eighteenth Century Collections Online* (ECCO); *English Short-Title Catalogue* (ESTC); metadata; Optical Character Recognition (OCR); eighteenth-century studies; bibliographic data science

1. INTRODUCTION

The relevance of quantitative-statistical methods for the description of the variation of English has increased rapidly during the last decades (cf. Gries 2012). In sync with the increase of the relevance of statistical or quantitative approaches to language, the availability of real-time language data, instead of tightly controlled corpora, has become a feature of corpus linguistics (Davies 2012). For historical studies of language change, the availability of data is the key question as to the basis of any work in the field (Hiltunen *et al.* 2017). However, creating a representative corpus is often difficult. Informal spoken language rarely survives (see, however, Hitchcock and Shoemaker 2007), letter collections are highly selective (already because of the question of literacy

DOI 10.32714/ricl.09.01.03

rates) and printed documents are biased towards higher classes of language users. Most large digitised collections also come with precious little information on the balance and biases within the corpus.

In relation to the eighteenth century, *Eighteenth Century Collections Online* (ECCO) has recently received attention not only from historians but from corpus linguists as well.[1] For example, the *Linguistic DNA* project aimed to use it as one of the main sources to uncover 'the DNA' of historical English discourse.[2] There are good reasons to take ECCO as the basis of studies on language variation. It is the most comprehensive dataset available in machine-readable form for eighteenth-century printed texts. It is linked to the *English Short-Title Catalogue* (ESTC)[3] that enables linking the collection to complementary text sources structured in the same way (most importantly *Early English Books Online* (EEBO),[4] which contains publications from 1473 to 1700). At the same time, it poses a series of problems. In the *Linguistic DNA* project, it was quickly realised that the quality of Optical Character Recognition (OCR) is highly problematic. Their conclusion was that "there are too many problems within the OCR dataset to use it" (Linguistic DNA 2017). One community-driven solution to these problems has been the *Text Creation Partnership*, which has turned to manual work to produce accurate transcriptions of a portion of the titles for EEBO and ECCO.[5] However, whereas for EEBO the EEBO-TCP collection covers almost half of the EEBO texts, ECCO-TCP contains transcriptions for only 3,101 out of the more than 200,000 texts in total. Therefore, as the OCRed version of ECCO is a remarkable source in size and scale, it is important to continue efforts towards making use of it in a reliable manner (Bullard 2013).

A systematic large-scale analysis of the biases in large digitised collections, such as ECCO and ESTC, can be critically complemented by algorithmic approaches (Lahti *et al.* 2015; Tolonen *et al.* 2018; Lahti *et al.* 2019; Lathi *et al.* 2020; Tolonen *et al.* 2021). Data quality is often suboptimal, posing challenges for large-scale comparisons

[1] ECCO ids referenced can be queried through the web-interface at https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online
[2] https://www.linguisticdna.org/
[3] The ESTC ids referenced can be queried through the web-interface of the British National Library at http://estc.bl.uk, and all the information regarding individual records is accessible through it. ESTC records used in this article have been enriched from the state of the version behind the web-interface implementation.
[4] !https://quod.lib.umich.edu/e/eebodemo/!
[5] https://textcreationpartnership.org/

and research use. The need for large-scale harmonisation has been widely recognised, and various solutions that are relevant to corpus linguistics are already available or have been proposed for the processing of digitised texts and other data types (Mäkelä *et al.* 2020). Overall, the applications of data science in this context aim at systematic and scalable improvements in data harmonisation, enrichment, and analysis, with the ultimate goal of advancing research on digital resources. Our present work relies heavily on our earlier efforts to harmonise the ESTC bibliographic metadata and the ongoing work to assess and potentially improve the quality of the ECCO full text collection. Here, we take the first steps towards a systematic integration and joint analysis of these two complementary sources. Whereas statistical integration of data from heterogeneous sources is a topical area in contemporary machine learning research, many pragmatic issues related to data quality and biases need to be understood and overcome before systematic and reliable statistical analyses can be carried out.

According to Davies (2012: 172) the main problems with large text archives (such as ECCO) are "accuracy, annotation, architecture, availability, and genre balance between different time periods." In this paper, we will look particularly at availability, architecture, genre balance and the accuracy in terms of OCR quality. We weigh these aspects of ECCO and its use in corpus linguistics from different perspectives and especially with respect to selection of corpora. If the magnitude of ECCO as big humanities data is seen as its best asset, how comprehensive is it in fact? We have harmonised the ESTC and worked connecting ECCO to the ESTC so that we can, for the first time, statistically evaluate the range of ECCO in the light of the ESTC.[6]

The aim of this paper is to reflect on different aspects of ECCO, in particular from the perspective of corpus linguistics. In Section 2.1, we give a statistical overview of ECCO in terms of different countries where works in ECCO were published and languages used in ECCO. We will then turn to discuss the temporal distribution of ECCO over the eighteenth century. In Section 2.2, a crucial part of our analysis is the analysis of reprints and new editions in ECCO (Ijaz *et al.* 2019) and, in Section 2.3, we will discuss the subject topics and genres in ECCO. After this, in Section 2.4, we turn to discuss the OCR quality of ECCO before concluding our observations in Section 3.

[6] We are currently writing a separate comprehensive article about the representativeness in ECCO when compared to ESTC.

2. ANALYSIS

ECCO was released in 2002 as a web-based query platform, after which it has been widely used at different universities by researchers and students alike. Originally, ECCO was scanned in the late 1990s from microfilms that date as far back as the early 1980s. Later in the 2000s, Gale —the company that owns the rights to distribute ECCO outside Britain— launched ECCO Part II (ECCO 2) that added 50,000 titles to the collection. In total, there are currently over 200,000 titles in the collection. Gale is at present digitising more materials with the intention to launch ECCO Part III with approximately 90,000 new titles later. Thus, it needs to be understood that already by its basic makeup ECCO is not a carefully selected or let alone balanced collection, but a layered historical source (about the history and development of ECCO including the selection process, see especially Gregg 2020. See also Kinley 2003; Greenfield 2010; Gale 2016; Cayley 2017).[7]

The more than 200,000 eighteenth-century documents included in ECCO amount to a little over 50 per cent of what is included in ESTC, the most comprehensive metadata collection of the British publication record for the early modern period (1470–1800). Thus, when compared to the publication record in general, ECCO is an impressive collection. There are however clear imbalances in the collection. In this article, we will discuss particularly geographical distribution, languages, temporal distribution, genre and estimates of OCR quality.[8] All our calculations are based on XML data dumps of ECCO Parts I and II obtained from Gale in 2015 through the Helsinki University Library, in accordance with Gale's updated text mining policy that allows researchers of a subscribing institution access to the content outside of Gale's user interface. All comparisons to ESTC are against our offline version graciously provided to us by the British Library in March 2016 and updated later.

2.1. Place, language and dating of publications

If we look at the geographical distribution of works in ECCO (Table 1), we quickly realise that especially items printed in the US are heavily underrepresented in the collection, compared most importantly to Scotland and Ireland. This bias can mainly be

[7] We are very grateful to Stephen Gregg for sharing his monograph with us prior to publication.
[8] We are also working on an analysis of different authors in the collection, but it is beyond the scope of this article.

explained by the origin of the digitised documents in ECCO, where the main part originate from the British Library and, to an important degree, also Oxford and Cambridge. While American libraries have also been part of the projects underlying ECCO, it is still clear that they remain heavily underrepresented in the dataset.

Country	**ESTC**	**ECCO**
England	233,473	134,935 (58%)
Scotland	33,864	17,365 (51%)
Ireland	24,957	16,647 (67%)
USA	40,672	10,088 (25%)
France	2,527	1,398 (55%)
Canada	995	35 (4%)
Others	4,517	2,157 (48%)
Unknown	2,868	1,133 (40%)
Total	**343,873**	**183,758 (53%)**

Table 1: Countries of publication in ECCO and the ESTC[9]

English is, by a vast margin, the dominant language in the nationally built collections of ESTC and ECCO (cf. Table 2). It is partly a reflection of ongoing changes in the British society at the time, especially since the number of Latin works is remarkably low compared to, for example, the eighteenth-century German and French sources (Lahti *et al.* 2019: 15–17). The presence of Welsh materials is noticeable in ECCO, while particularly German sources are missing. Within English language publications, what is important for the study of language variation is that the number of publications in both Ireland and Scotland is high. Even when most of the Dublin printing activity focused on London reprints, there is still a good chance to use these materials to identify regional variation in language use in Britain. We consider this as one of the prominent research fields with respect to ECCO.

[9] The number of ESTC records for the same time period (1701–1800) is shown for comparison. The percentages indicate the fraction ESTC records that are coverd by ECCO. The aggregate 'Others' includes a mixed bag of all countries with fewer records in these collections than Canada, such as Belgium, Germany, Italy, Switzerland, and the Netherlands, but also Barbados, Haiti, India, Jamaica, and so forth. The Category 'Unknown' consists of records whose place of publication is recorded.

Primary language	ESTC	ECCO
English	324,804	173,967 (54%)
Latin	7,699	4,599 (60%)
French	7,269	3,783 (52%)
Welsh	765	540 (71%)
Italian	510	341 (67%)
German	1,630	279 (17%)
Others	1,196	249 (21%)
Total	**343,873**	**183,758 (53%)**

Table 2: Main languages in ECCO and the ESTC[10]

One aspect that needs to be taken into consideration when using ECCO for text mining is that the corpus is uneven over time. From 1780 to 1800 there are far more documents than during the earlier decades (cf. Figure 1). Since there are also changes in the distribution of genres during this time, this obviously is something that needs to be taken into consideration when using ECCO as a corpus.

Figure 1: Variation in ECCO title count during the 18th century

Another important point is that some of the dates in ECCO are uncertain. Thus, a document dated for a particular year (particularly even years such as 1710) might actually be from any year during that decade. In many cases, the uncertainty has been indicated in the ESTC (with e.g. a question mark, 'ca.', or time range), and we have

[10] The aggregate "Others" includes Ancient Greek, Dutch, Hebrew, Portuguese, Spanish, Tamil, and a number of other languages.

used this to identify the uncertain years (shown in purple in Figure 1.). These uncertain attributions contribute peaks to even five, ten and 50 years.

2.2. Reprints

A further aspect that anyone using ECCO as a corpus needs to take into account is that a large part of the collection are reprints and further editions of previously published works. Gale, in their online materials, has suggested that new editions should contain substantial new material in order to be included, and that mere reprints would be for the most part excluded.[11] Based on our evaluation, however, this is not true and some titles are repeated dozens of times, years after their initial publication, while others are missing from the collection altogether.[12] This obviously has quite an impact on the general shifts in language that we might detect from the collection. One way of phrasing this is that we may get two different perspectives to language when using ECCO. If we use the collection as a whole, our perspective is the language available to readers at a particular time. Here it is obvious that if a particular work is printed verbatim several times over, it has more impact molding the minds of the reading public. We may look at the classics, for example, from this perspective. The other viewpoint would be to make a subset of ECCO that would include only any possibly novel parts of later editions past the first publication. If we are interested in neologisms, for example, this might be a more viable approach, because the dataset would only include new works and thus tracking the emergence and diffusion of new language might be easier.

With respect to duplication, two distinct viewpoints can be considered. First, we consider duplication within ECCO itself. Based on our analysis, in the 184,029 ESTC records contained in ECCO, there are 115,962 unique works. Therefore, a full 37 per cent of the content within ECCO may be duplicated elsewhere within it. Of the distinct works, 80 per cent appear inside ECCO only once, 11 per cent twice and nine per cent more than two times, with Thomas Sternhold and John Hopkins' *Book of Psalms* holding top place with 135 copies, followed by John Milton's *Paradise Lost* with 118.

---

[11] Originally, the *Eighteenth Century* microfilm project was limited to "first and significant editions of each title" with the exception of 28 major authors whose editions were all included (Alston 1981: 2). This is still visible in ECCO. For the full history of the complicated selection process behind ECCO, see Gregg (2020).

[12] For our process of identifying reprints, see Ijaz *et al.* (2019).

As a second viewpoint, we consider the amount of material in ECCO that are reprints from earlier years, without regard to whether the original versions are included in ECCO themselves. Via this viewpoint, we can consider how well the texts associated with a particular year in ECCO actually correspond to contemporary language, as opposed to the language of years past. As we notice in Figure 2, the fraction of material that are reprints from earlier years in ECCO grows particularly towards the 1750s up to >30 per cent.[13] In total, a full 31 per cent of the titles in ECCO are reprints of some kind, highlighting the increasing importance of reprints among the overall publishing activities. Naturally, new editions contain some new language (and, in some cases, also extensive additions to the original work), but eighteenth-century printing technology favoured exact reprinting (cf. Bonnell 2009). In terms of evaluating the bias caused by these reprints, it is interesting to know their age distribution. Here, the median age of the reprints is seventeen years from their first printing, but the spread is large, with a full nine per cent of reprints dating back more than 100 years ago.

Figure 2: Share of ECCO that are reprints from earlier years for the period 1701–1800

More bias is added to this equation when we realise that the presence of popular authors in ECCO is prominent. This seems partly to be a legacy of the *Eighteenth Century* microfilm project where a decision was made to include all the editions of the works of

[13] The graph identifies the percentage of reprints each year as identified from ESTC. For our method of detecting reprints, see Ijaz *et al*. (2019). Also, texts printed before 1700 are included in the graph.

twenty eight authors considered 'major'.[14] As a result, the editions of, for example, Henry Fielding, Alexander Pope, Samuel Johnson and Laurence Sterne are nearly completely covered in ECCO, while the works of other eighteenth-century authors are not included, let alone all the editions of these works. This is a serious form of bias in the collection because it amplifies the effect of the already well-known and studied authors. Thus, we need to be careful not to take the language of Pope and Johnson, for instance, to represent the eighteenth century in general because of imbalances in the corpus that we study.

2.3. Subject headings

One feature of ECCO is that it includes subject headings for all the documents. This is also a legacy of the *Eighteenth Century* microfilm project where the collection was arranged to eight subject heading categories.[15]

When we examine the subject heading distribution over time, we realise that there are both lasting trends as well as spot anomalies in the data. Taking the proportion of running words in each section as a measure (cf. Figure 3 below), we see first of all that the share of 'Religion and Philosophy' goes down over the eighteenth century, whereas the role of 'Literature and Language' grows somewhat over time. At the same time, there is a significant anomaly in the 1730s where the proportion of words associated with 'General Reference' suddenly spikes upwards. Upon investigation, this spike is caused solely by the inclusion in ECCO of two separate 1734 editions of Pierre Bailey's dictionary, consisting of five and ten volumes of around a thousand pages each. Given that the language of such dictionaries is certainly a distinct genre with more precise definitions of words and concepts, not filtering these out may certainly affect any text mining results based on ECCO. Earlier we have examined this aspect with respect to use of philosophical language, and it turns out that towards the later

[14] Addison, Bentham, Bishop Berkeley, Boswell, Burke, Burns, Congreve, Defoe, Jonathan Edwards, Fielding, Franklin, Garrick, Gibbon, Goldsmith, Hume, Johnson, Paine, Pope, Reynolds, Richardson, Bolingbroke, Sheridan, Adam Smith, Smollett, Steele, Sterne, Swift and Wesley. For further discussion, see Gregg (2020).

[15] According to Gregg (2020: 21), "these subject headings may well have had their origin in Alston's experiments with the 18thC STC's initial online interface at the British Library, which he felt could help in the creation of subject packages which will form the basis of the RPI program to microfilm the substantive texts in ESTC (Alston 2004)."

eighteenth century the growth in precise definitions of philosophical concepts is considerable (Tolonen *et al.* 2017).

Figure 3: Composition of ECCO 1 in terms of the number of words by subject heading and year[16]

Apart from the few very large dictionaries and the anomalies they cause, the 'General Reference' and 'Law' categories, on the other hand, are much smaller in ECCO than they are in reality. This is because it is especially the almanacs, proclamations, general acts and the like that were intentionally excluded from the materials that form ECCO (Alston 1981). Yet also here there is a temporal anomaly. For reasons unknown to us, from the 1750s to the 1770s, a much larger amount of bills and petitions has been included. Due to these being very short, this anomaly is mostly not discernible in the 'Law' data of Figure 3 but does show up clearly if the data is weighted by the number of publications instead of the number of words in them.

2.4. OCR quality

As ECCO is a corpus arising from automated mass digitisation, it is susceptible to noise from the OCR process. In earlier work (cf. Hill and Hengchen 2019) comparing the ECCO-TCP hand-transcribed subset of ECCO 1 to the OCRed version, it was identified

[16] The integration of multi-volume titles and their impact on the numerical estimates are influenced by variations in publication years, edition counts, and other factors. A full manual curation of the large data collection is here replaced by an approximation, where the multiple volumes are aggregated and counted at the first occurrence. This makes it possible to scale up the estimates to cover the whole data collection but may introduce additional bias, such as the peak that we can observe at Bailey's 1734 dictionary.

that, on an overall level, the token-level mean precision of ECCO OCR is 0.744 (meaning that on average, 74% of the tokens in ECCO OCR are correct), with recall being 0.814 (meaning that 81% of the tokens in the original are included in the OCRed version).

While the above results speak directly only for the small ECCO-TCP subset, we also identified a statistically significant ($p<0.001$) Pearson correlation of 0.795 between the page-level F1 score (the harmonic mean of precision and recall) and the confidence value reported by the OCR engine used by Gale. This agreement supports being able to use the OCR engine confidence value to accurately assess OCR quality also beyond the small subset.

However, ECCO 1 and ECCO 2 arise from different OCR processes, and the above correlation strictly applies only to ECCO 1 due to the ECCO-TCP only containing material from it. Yet, the confidence scores for both ECCO 1 and ECCO 2 do follow similar patterns with regard to time, language and other secondary axes, suggesting that also the ECCO 2 engine confidence could be trusted.

Importantly though, the confidence estimates of the OCR engine used for digitising ECCO 2 are probably not directly comparable to those reported by the engine used for ECCO 1. To wit, the confidence scores reported by the ECCO 2 process are consistently lower than those reported by the ECCO 1 process. Instead of indicating a general decrease in OCR quality for the publications scanned later, this more likely just means that the confidence estimates operate on different scales overall.

Figure 4 charts the OCR confidence measures in the two subcollections against time. For both collections, median accuracy improves with time, particularly from 1700 to 1750. At the same time, both collections contain many outliers with a remarkably lower confidence. On a surface level, one would also be tempted to draw the conclusion that the quality variation is more intense in ECCO 2, but that may just be an artifact of the different confidence scales used in the two collections.

Figure 4: OCR quality in ECCO 1 and ECCO 2 through time[17]

While ECCO is primarily composed of English texts, if one is interested in the small subparts of it which are not, one will be interested in how the OCR quality is affected by the language. First, to verify whether language had an effect on the reliability of the OCR confidence estimates, we calculated the correlation between ECCO-TCP transcriptions and ECCO OCR versions for the different languages. That collection contains only a few documents in languages other than English, including French (N=31) and Welsh (N=94), and 443 documents with an unknown language. The correlation between the manually curated quality (F1 scores) and the automated OCR confidence intervals was 0.8 across all languages without any significant difference. Thus, the confidence scores seem to be trustworthy indicators of OCR quality also for non-English documents.

Expanding from this to look at the OCR confidences across all languages in ECCO (cf. Figure 5), we see that the median confidence is lower for languages other than English. Of particular interest here is that German has a remarkably lower OCR confidence than the other languages. This might be due to the system not being properly configured for German special characters, which do appear in the automated transcriptions, but not as often as they should.

[17] Note that the OCR confidences provided by the engine (vertical axis) are not comparable between ECCO 1 and ECCO 2.

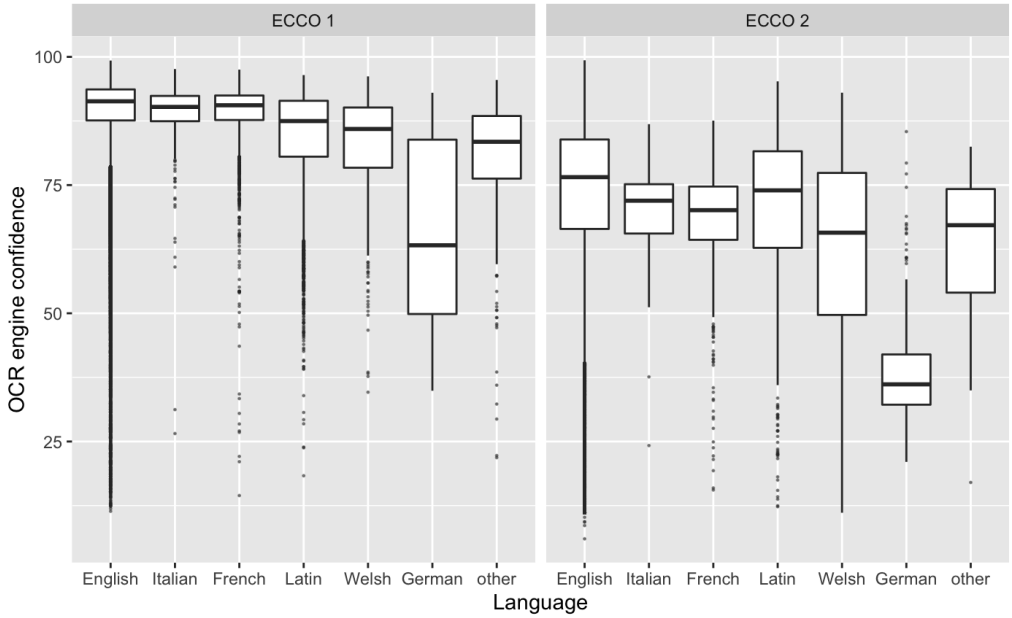

Figure 5: OCR quality in different languages[18]

3. Conclusion

ECCO is a primary source for anyone interested in eighteenth-century English language. The availability of ECCO for text mining is also changing the way scholars work (and will work in the future). While these kinds of big data sources will gain even more prominence in the future, the role of source criticism will be more and more important to all fields that want to use large historical collections. The interests of linguists, historians and data scientists are thus mutual and all these relevant expertises are needed.

Our analysis of ECCO has shown that different kinds of biases in the data are evident based on the general composition of the collection alone. The geographical distribution of ECCO is uneven compared to the full eighteenth-century British printing record. There are also historical reasons for the geographical imbalance but the main reason for the missing documents in English is the process of putting ECCO together. The temporal distribution of ECCO is likewise uneven, with the end of the eighteenth century dominating the corpus mainly because the printing activity was increasing during that time. The reprint activity, too, is higher towards the end of the century, and there are more reprints included during the later eighteenth-century decades in ECCO than earlier. It is also evident that the most popular authors are overrepresented, which

[18] Note that the OCR quality scores (vertical axis) are not comparable between ECCO 1 and ECCO 2.

creates a bias of its own in the reprints that can be detected in the data. The OCR quality in ECCO data is generally remarkably lower compared to results that are achieved in digitisation of scanned sources in the 2020s. Furthermore, the OCR quality is significantly uneven between different parts of ECCO. When we combine this information about the OCR quality with the question of reprints and other issues discussed in our analysis we understand that these biases accumulate. This is visible for example in basic key-word searches. Popular authors are overrepresented already for historical reasons and their presence in ECCO is further amplified due to other biases in the selection process of included works and poor OCR quality. Obviously, the more works you have included in the data, the greater the likelihood of them turning up on different occasions.

There are good reasons why we should take the opportunity to use ECCO when studying language change seriously. ECCO is a remarkable source in spite of the gaps in the data that we have detected. When we combine an understanding of possible bias in the data with the potential of ECCO for data mining, we may formulate more robust approaches to it in our research. There is great potential in ECCO to study language variation and change when we take into consideration the distinction between ‘a corpus as the input for a reader’ (canonical works or ideas) and ‘a corpus as the output of a writer’ (neologisms), the increase in precise definitions of philosophical concepts, and the correlation between OCR engine confidence and quality. For example, the regional variation of eighteenth-century printed English is an aspect that we can study based on this source. But what is needed is the understanding of the historicity of the source both as actual historical processes and also as the layering of a collection that has a complicated provenance. After grasping this historicity, we are then able to think of different ways to limit the effect of these biases.

We believe that investigating language by the use of ECCO is possible, given that careful work is put into taking different aspects into consideration and the research questions are matched with what is possible to do with such a biased and largely inaccurate corpus. Our aim in this article has been to bring forward some crucial limitations of ECCO in the use of corpus linguistics. The next step will be to overcome these limitations, especially with respect to the low OCR quality that renders many

intuitively useful interfaces for modelling ECCO, such as Gale's own *Digital Scholar Lab*, currently virtually unusable for many research tasks.[19]

References

Alston, Robin. 1981. ESTC texts on microfilm. *Factotum: Newsletter of the XVIIIth century STC* 12: 2–3.

Alston, Robin. 2004. The history of ESTC. *The Age of Johnson* 15: 269–329.

Bonnell, Thomas F. 2009. Reprint trade. In Michael F. Suarez and Michael L. Turner eds. *The Cambridge History of the Book in Britain. Vol. V. 1695–1830*. Cambridge: Cambridge University Press, 699–709.

Bullard, Paddy. 2013. Digital humanities and electronic resources in the long eighteenth century. *Literature Compass* 10/10: 748–760.

Cayley, Seth. 2017. Digitization for the masses: Taking users beyond simple searching in Nineteenth-Century Collections Online. *Journal of Victorian Culture* 22/2: 248–255.

Davies, Mark. 2012. Some methodological issues related to corpus-based investigations of recent syntactic changes in English. In Terttu Nevalainen and Elizabeth Closs Traugott eds. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 157–174.

Eighteenth Century Collections Online. https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online

English Short Title Catalogue. http://estc.bl.uk

Gale. 2016. *Eighteenth Century Collections Online*: The most comprehensive online library of English and foreign-language titles printed in the United Kingdom during the eighteenth century, plus thousands of important works printed in English elsewhere. https://www.gale.com/binaries/content/assets/gale-us-en/primary-sources/eighteenth-century-collections-online/ecco-roll-fold-2016-web.pdf

Greenfield, Sayre. 2010. ECCO OCR troubleshooting. *Early Modern Online Bibliography*. https://earlymodernonlinebib.wordpress.com/ecco-ocr-troubleshooting-by-sayre-greenfield/ (15 January, 2020.)

Gregg, Stephen. 2020. *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2012. Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: Towards more and more fruitful exchanges. In Joybrato Mukherjee and Magnus Huber eds. *Corpus Linguistics and Variation in English. Theory and Description*. Amsterdam: Rodopi, 41–63.

Hill, Mark J. and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: *Eighteenth Century Collections Online* as a case study. *Digital Scholarship in the Humanities* 34/4: 825–843.

Hiltunen, Turo, Joe McVeigh and Tanja Säily. 2017. How to turn linguistic data into evidence? In Turo Hiltunen, Joe McVeigh and Tanja Säily eds. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations*. Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/19/introduction.html (24 April, 2021.)

[19] https://www.gale.com/intl/primary-sources/digital-scholar-lab

Hitchcock, Tim and Robert Shoemaker. 2007. The value of the proceedings as a historical source. *Old Bailey Proceedings Online*. https://www.oldbaileyonline.org/static/Value.jsp (16 April, 2021.)

Ijaz, Ali, Leo Lahti, Iiro Tiihonen and Mikko Tolonen. 2019. Analytical determination of editions from bibliographic metadata. In Jarmo Harri Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen and Katja Västi eds. *Proceedings of the Research Data and Humanities 2019 Conference: Data, Methods and Tools*. Oulu: University of Oulu. http://urn.fi/urn:isbn:9789526223216 (24 April, 2021.)

Kinley, Welly. 2003. Digital ECCOs of the eighteenth century. *eContent*, November Issue. https://chnm.gmu.edu/digitalhistory/links/pdf/introduction/0.27b.pdf (24 April, 2021.)

Lahti, Leo, Niko Ilomäki and Mikko Tolonen. 2015. A quantitative study of history in the *English Short-Title Catalogue* (ESTC) 1470–1800. *LIBER Quarterly* 25/2: 87–116.

Lahti Leo, Eetu Mäkelä and Mikko Tolonen. 2020. Quantifying bias and uncertainty in historical data collections with probabilistic programming. In Folgert Karsdorp, Barbara McGillivray, Adina Nerghes and Melvin Wevers eds. *Proceedings of the Workshop on Computational Humanities Research 2020.* Aachen: CEUR-WS.org, 280–289.

Lahti, Leo, Jani Marjanen, Hege Roivainen and Mikko Tolonen. 2019. Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly* 57/1: 5–23.

Linguistic DNA. 2017. Experimenting with the imperfect: ECCO & OCR. https://www.linguisticdna.org/ecco-ocr/ (20 February, 2020.)

Mäkelä, Eetu, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi and Terttu Nevalainen. 2020. Wrangling with non-standard data. In Sanita Reinsone, Inguna Skadiņa, Anda Baklāne and Jānis Daugavietis eds. *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference 2020*. Aachen: CEUR-WS.org, 81–96.

Tolonen, Mikko, Eetu Mäkelä and Leo Lahti. 2017. Analysing eighteenth-century key-terms and phrases using ECCO and ESTC. *Paper presented at the British Society for Eighteenth Century Studies BSECS 46th Annual Conference*, Oxford.

Tolonen, Mikko, Leo Lahti, Jani Marjanen and Hege Roivainen. 2018. A quantitative approach to book-printing in Sweden and Finland, 1640–1828. *Historical Methods* 52/1: 57–78.

Tolonen Mikko, Mark Hill, Ali Ijaz, Ville Vaara and Leo Lahti. 2021. Examining the early modern canon: The *English Short Title Catalogue* and large-scale patterns of cultural production. In Ileana Baird ed. *Data Visualization in Enlightenment Literature and Culture*. London: Palgrave Macmillan, 63–119.

Corresponding author
Mikko Tolonen
University of Helsinki
P.O. Box 24
00014. Helsinki
Finland
Email: mikko.tolonen@helsinki.fi

received: March 2020
accepted: April 2021